New web technology would let you track how your private data is used online

June 13 2014, by Larry Hardesty

(Left to right) Tim Berners-Lee, Oshani Seneviratne, and Lalana Kagal. Credit: Bryce Vickmark

By now, most people feel comfortable conducting financial transactions on the Web. The cryptographic schemes that protect online banking and credit card purchases have proven their reliability over decades.

As more of our data moves online, a more pressing concern may be its

inadvertent misuse by people authorized to access it. Every month seems to bring another story of private information accidentally leaked by governmental agencies or vendors of digital products or services.

At the same time, tighter restrictions on access could undermine the whole point of sharing data. Coordination across agencies and providers could be the key to quality medical care; you may want your family to be able to share the pictures you post on a social-networking site.

Researchers in the Decentralized Information Group (DIG) at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) believe the solution may be transparency rather than obscurity. To that end, they're developing a protocol they call "HTTP with Accountability," or HTTPA, which will automatically monitor the transmission of private data and allow the data owner to examine how it's being used.

At the IEEE's Conference on Privacy, Security and Trust in July, Oshani Seneviratne, an MIT graduate student in electrical engineering and computer science, and Lalana Kagal, a principal research scientist at CSAIL, will present a paper that gives an overview of HTTPA and presents a sample application, involving a health-care records system that Seneviratne implemented on the experimental network PlanetLab.

DIG is directed by Tim Berners-Lee, the inventor of the Web and the 3Com Founders Professor of Engineering at MIT, and it shares office space with the World Wide Web Consortium (W3C), the organization, also led by Berners-Lee, that oversees the development of Web protocols like HTTP, XML, and CSS. DIG's role is to develop new technologies that exploit those protocols.

With HTTPA, each item of private data would be assigned its own uniform resource identifier (URI), a key component of the Semantic Web, a new set of technologies, championed by W3C, that would

convert the Web from, essentially, a collection of searchable text files into a giant database.

Remote access to a Web server would be controlled much the way it is now, through passwords and encryption. But every time the server transmitted a piece of sensitive data, it would also send a description of the restrictions on the data's use. And it would log the transaction, using only the URI, somewhere in a network of encrypted, special-purpose servers.

HTTPA would be voluntary: It would be up to software developers to adhere to its specifications when designing their systems. But HTTPA compliance could become a selling point for companies offering services that handle private data.

"It's not that difficult to transform an existing website into an HTTPA-aware website," Seneviratne says. "On every HTTP request, the server should say, 'OK, here are the usage restrictions for this resource,' and log the transaction in the network of special-purpose servers."

An HTTPA-compliant program also incurs certain responsibilities if it reuses data supplied by another HTTPA-compliant source. Suppose, for instance, that a consulting specialist in a network of physicians wishes to access data created by a patient's primary-care physician, and suppose that she wishes to augment the data with her own notes. Her system would then create its own record, with its own URI. But using standard Semantic Web techniques, it would mark that record as "derived" from the PCP's record and label it with the same usage restrictions.

The network of servers is where the heavy lifting happens. When the data owner requests an audit, the servers work through the chain of derivations, identifying all the people who have accessed the data, and what they've done with it.

Seneviratne uses a technology known as distributed hash tables—the technology at the heart of peer-to-peer networks like BitTorrent—to distribute the transaction logs among the servers. Redundant storage of the same data on multiple servers serves two purposes: First, it ensures that if some servers go down, data will remain accessible. And second, it provides a way to determine whether anyone has tried to tamper with the transaction logs for a particular data item—such as to delete the record of an illicit use. A server whose logs differ from those of its peers would be easy to ferret out.

To test the system, Seneviratne built a rudimentary health-care records system from scratch and filled it with data supplied by 25 volunteers. She then simulated a set of transactions—pharmacy visits, referrals to specialists, use of anonymized data for research purposes, and the like—that the volunteers reported as having occurred over the course of a year.

Seneviratne used 300 servers on PlanetLab to store the transaction logs; in experiments, the system efficiently tracked down data stored across the network and handled the chains of inference necessary to audit the propagation of data across multiple providers. In practice, audit servers could be maintained by a grassroots network, much like the servers that host BitTorrent files or log Bitcoin transactions.

More information: The paper, "Enabling Privacy Through Transparency," is available online: dig.csail.mit.edu/2014/Papers/PST-PETS/PETS.pdf

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology